

**METHOD AND SYSTEM FOR THE AUTOMATIC AMENDMENT OF SPEECH  
RECOGNITION VOCABULARIES**

Inventors

Werner Kriechbaum

Gerhard Stenzel

International Business Machines Corporation

IBM Docket No. DE9-2000-0096

IBM Disclosure No. DE8-2000-0115

EXPRESS MAILING LABEL NO.: EL 740159505 US

**CROSS-REFERENCE TO RELATED APPLICATIONS**

This application claims the benefit of European Application No. 00127484.4, filed November 29, 2000 at the European Patent Office.

5

**BACKGROUND OF THE INVENTION**

**Technical Field**

The invention generally relates to the field of computer-assisted or computer-based speech recognition, and more specifically, to a method and system for improving recognition quality of a speech recognition system.

10

**Description of the Related Art**

Conventional speech recognition systems (SRSs), in a very simplified view, can include a database of word pronunciations linked with word spellings. Other supplementary mechanisms can be used to exploit relevant features of a language and the context of an utterance. These mechanisms can make a transcription more robust. Such elaborate mechanisms, however, will not prevent a SRS from failing to accurately recognize a spoken word when the database of words does not contain the word, or when a speaker's pronunciation of the word does not agree with the pronunciation entry in the database. Therefore, collecting and extending vocabularies is of prime importance for the improvement of SRSs.

Presently, vocabularies for SRSs are based on the analysis of large corpora of written documents. For languages where the correspondence between written and spoken language is not bijective, pronunciations have to be entered manually. This is a laborious and costly procedure.

U.S. Patent No. 6,064,957 discloses a mechanism for improving speech recognition through text-based linguistic post-processing. Text data generated from a SRS and a corresponding true transcript of the speech recognition text data are collected and aligned by means of a text aligner. From the differences in alignment, a

15  
20

plurality of correction rules are generated by means of a rule generator coupled to the text aligner. The correction rules are then applied by a rule administrator to new text data generated from the SRS. The mechanism performs only a text-to-text alignment, and thus does not take the particular pronunciation of the spoken text into account.

5 Accordingly, it needs the aforementioned rule administrator to apply the rules to new text data. The mechanism therefore cannot be executed fully automatically.

U.S. Patent No. 6,078,885 discloses a technique which provides for verbal dictionary updates by end-users of the SRS. In particular, a user can revise the phonetic transcription of words in a phonetic dictionary, or add transcriptions for words 10 not present in the dictionary. The method determines the phonetic transcription based on the word's spelling and the recorded preferred pronunciation, and updates the dictionary accordingly. Recognition performance is improved through the use of the updated dictionary.

The above discussed techniques, however, share the disadvantage of not being able to update a speech recognition vocabulary on large scale bodies of text with minimal technical effort and time. Accordingly, these techniques are not fully automated.

1500 1400 1200 1000 800 600 500 400 300 200 100

### SUMMARY OF THE INVENTION

It is therefore an object of the present invention to provide method and system for improving the recognition quality and quantity of a speech recognition system. It is another object to provide such a method and system which can be executed or 5 performed automatically. Another object is to provide a method and system for improving the recognition quality with minimum technical effort and time. It is yet another object to provide such a method and system for processing large text corpora for updating a speech recognition vocabulary.

The above objects are solved by the features of the independent claims. Other 10 advantageous embodiments are disclosed within the dependent claims. Speech recognition can be performed on an audio realization of a spoken text to derive a hypothesis textual representation (second representation) of the audio realization. Using the recognition results, the second representation can be compared with an allegedly true textual representation (first representation), i.e. an allegedly correct transcription of the audio realization in a text format, to look for non-recognized single 15 words. These single words then can be used to update a user-dictionary (vocabulary) or pronunciation data obtained by a training of the speech recognition.

It is noted that the true textual representation (true transcript) can be obtained in a digitized format, e.g. using known character recognition (OCR) technology. Further it 20 has been recognized that an automation of the above mentioned mechanism can be achieved by providing a looped procedure where the entire audio realization and both the entire true textual representation and the speech-recognized hypothesis textual representation can be aligned to each other. Accordingly, the true textual representation and the hypothetic textual representation likewise can be aligned to each 25 other. The required information concerning mis-recognized or non-recognized speech segments therefore can be used together with the alignment results in order to locate mis-recognized or non-recognized single words.

Notably, the proposed procedure of identifying isolated mis-recognized or non-recognized words in the entire realization and representation, and to correlate

these words in the audio realization, advantageously makes use of an inheritance of the time information from the audio realization and the speech recognized second transcript to the true transcript. Thus, the audio signal and both transcriptions can be used to update a word database, a pronunciation database, or both.

5 The invention disclosed herein provides an automated vocabulary or dictionary update process. Accordingly, the invention can reduce the costs of vocabulary generation, e.g. of novel vocabulary domains. The adaptation of a speech recognition system to the idiosyncrasies of a specific speaker is currently an interactive process where the speaker has to correct mis-recognized words. The invention disclosed  
10 herein also can provide an automated technique for adapting a speech recognition system to a particular speaker.

15 The invention disclosed herein can provide a method and system for processing large audio or text files. Advantageously, the invention can be used with an average speaker to automatically generate complete vocabularies from the ground up or generate completely new vocabulary domains to extend an existing vocabulary of a speech recognition system.

**BRIEF DESCRIPTION OF THE DRAWINGS**

There are shown in the drawings embodiments which are presently preferred, it being understood, however, that the invention is not limited to the precise arrangements and instrumentalities shown.

5 Fig. 1 is a block diagram illustrating a system in accordance with the inventive arrangements disclosed herein.

Fig. 2 is a block diagram of an aligner configured to align a true textual representation and a hypothesis timed transcript in accordance with the inventive arrangements disclosed herein.

10 Fig. 3 is a block diagram of a classifier configured to process the output of the aligner of Fig. 2 in accordance with the inventive arrangements disclosed herein.

Fig. 4 is a block diagram illustrating inheritance of timing information in a system in accordance with the inventive arrangements disclosed herein.

15 Fig. 5 is an exemplary data set consisting of a true transcript, a hypothesis transcript provided through speech recognition, and a corresponding timing information output from an aligner in accordance with the inventive arrangements disclosed herein.

Fig. 6 depicts an exemplary data set output from a classifier in accordance with the inventive arrangements disclosed herein.

20 Fig. 7 illustrates corresponding data in accordance with a first embodiment of the inventive arrangements disclosed herein.

Fig. 8 illustrates corresponding data in accordance with a second embodiment of the inventive arrangements disclosed herein.

**DETAILED DESCRIPTION OF THE INVENTION**

Fig. 1 provides an overview of a system and a related procedure in accordance with the inventive arrangements disclosed herein by way of a block diagram. The procedure starts with a realization 10, preferably an audio recording of human speech, i.e. a spoken text, and a representation 20, preferably a transcription of the spoken text.

5 Many pairs of an audio realization and a true transcript (resulting from a correct transcription) are publically available, e.g. radio features stored on a storage media such as CD-ROM and the corresponding scripts, or audio versions of text books primarily intended for teaching blind people.

10 The realization 10 is first input to a speech recognition engine 50. The textual output of the speech recognition engine 50 and the representation 20 are aligned by means of an aligner 30. The aligner 30 is described in greater detail with reference to Fig. 2. The output of the aligner 30 is passed through a classifier 40. The classifier 40 is described in greater detail with reference to Fig. 3. The classifier compares the aligned representation with a transcript produced by the speech recognition engine 50 and tags all isolated single word recognition errors. An exemplary data set is depicted in Fig. 5.

15 In a first embodiment of the present invention, a selector 60 can select all one word pairs for which the representation and the transcript are different (see also Fig. 6). The selected words, together with their corresponding audio signal, are then used to update a word database. In a second embodiment, word pairs for which the representation and the transcript are similar, are selected for further processing. The selected words, together with their corresponding audio signal, are then used in the second embodiment to update a pronunciation database of a speech recognition system.

20

25 Referring to Fig. 2, an aligner can be used by the present invention to align a true representation 100 and a hypothesis timed transcript 110. In a first step 120, acronyms and abbreviations can be expanded. For example, short forms like 'Mr.' are expanded to the form 'mister' as they are spoken. In a second step all markup is

stripped 130 from the text. For plain ASCII texts, this procedure removes all punctuation marks such as “;”, “,”, “.”, and the like. For texts structured with a markup language, all the tags used by the markup language can be removed. Special care can be taken in cases where the transcript has been generated by a SRS system, as is the case in the method and system according to the present invention working in dictation mode. In this case, the SRS system relies on a command vocabulary to insert punctuation marks which have to be expanded to the words used in the command vocabulary. For example, “.” is replaced by “full stop”.

After both texts, the time-tagged transcript generated by the SRS and the representation, have been “cleaned” or processed as described above, an optimal word alignment 140 is computed using state-of-the-art techniques as described in, for example, Dan Gusfield, “Algorithms on Strings, Trees, and Sequences”, Cambridge University Press Cambridge (1997). The output of this step is illustrated in Fig. 5 and includes 4 columns. For each line, 600 gives the segments of the representation that aligns with the segment of the transcript 610. 620 provides the start time and 630 provides the end time of the audio signal that resulted in the transcript 610. It should be noted that due to speech recognition errors the alignment between 610 and 620 is not 1-1 but m-n, i.e. m words of the realization may be aligned with n words of the transcript.

Fig. 3 is an overview block diagram of the classifier that processes the output of the aligner described above. For all lines 200 in Fig. 5, the classifier adds 210 an additional entry in column 740 as shown in Fig. 6. The entry specifies whether the correspondence between the representation and the transcript is 1-1. For each line of the aligner output, the classifier tests 220 whether the entry consists of one word. If this is not true, the value ‘0’ is added 240 in column 740 and the next line of the aligner output is processed. If the entry in column 700 consists only of one word, the same test 230 is applied to the entry in column 710. If this entry also consists only of one word, the value ‘1’ is added 250 in column 740. Otherwise the value ‘0’ is written in 740.

Fig. 4 is a block diagram illustrating the inheritance of timing information in a system in accordance with the inventive arrangements disclosed herein. An audio realization, in the present embodiment, is input real-time to a SRS 500 via microphone 510. Alternatively, the audio realization can be provided offline together with a true transcript 520 which already has been checked for correctness of the assumed preceding transcription process. It is further assumed that the SRS 500 reveals a timing information for the audio realization. Thus, the output of the SRS 500 is a potentially correct transcript 530 which includes timing information and the timing information 540 itself which can be accessed separately from the recognized transcript 530.

10

15

20

25

30

35

40

45

50

55

60

65

70

75

80

85

90

95

100

105

110

115

120

125

130

135

140

145

150

155

160

165

170

175

180

185

190

195

200

205

210

215

220

225

230

235

240

245

250

255

260

265

270

275

280

285

290

295

300

305

310

315

320

325

330

335

340

345

350

355

360

365

370

375

380

385

390

395

400

405

410

415

420

425

430

435

440

445

450

455

460

465

470

475

480

485

490

495

500

505

510

515

520

525

530

535

540

545

550

555

560

565

570

575

580

585

590

595

600

605

610

615

620

625

630

635

640

645

650

655

660

665

670

675

680

685

690

695

700

705

710

715

720

725

730

735

740

745

750

755

760

765

770

775

780

785

790

795

800

805

810

815

820

825

830

835

840

845

850

855

860

865

870

875

880

885

890

895

900

905

910

915

920

925

930

935

940

945

950

955

960

965

970

975

980

985

990

995

1000

1005

1010

1015

1020

1025

1030

1035

1040

1045

1050

1055

1060

1065

1070

1075

1080

1085

1090

1095

1100

1105

1110

1115

1120

1125

1130

1135

1140

1145

1150

1155

1160

1165

1170

1175

1180

1185

1190

1195

1200

1205

1210

1215

1220

1225

1230

1235

1240

1245

1250

1255

1260

1265

1270

1275

1280

1285

1290

1295

1300

1305

1310

1315

1320

1325

1330

1335

1340

1345

1350

1355

1360

1365

1370

1375

1380

1385

1390

1395

1400

1405

1410

1415

1420

1425

1430

1435

1440

1445

1450

1455

1460

1465

1470

1475

1480

1485

1490

1495

1500

1505

1510

1515

1520

1525

1530

1535

1540

1545

1550

1555

1560

1565

1570

1575

1580

1585

1590

1595

1600

1605

1610

1615

1620

1625

1630

1635

1640

1645

1650

1655

1660

1665

1670

1675

1680

1685

1690

1695

1700

1705

1710

1715

1720

1725

1730

1735

174

medical treatment field can be added. The proposed mechanism selects lines of the output of the classifier (Fig. 7) which include a tag bit of "1", but include only non-  
identical single words such as "Wahn" and "Mann" in the present example. These  
single words represent single word recognition errors of the underlying speech  
recognition engine, and therefore can be used in a separate step to update a word  
database of the underlying SRS.

5 A second embodiment of the present invention, as illustrated in Fig. 8, provides  
for an automated speaker related adaptation of an existing vocabulary which does not  
require active training through the speaker. Accordingly, only single words where the  
10 tag bit equals "1" are selected for which the true transcript (left column) and the  
recognized transcript (right column) are identical (Fig. 8). These single words represent  
correctly recognized isolated words and thus can be used in a separate step to update  
a pronunciation database of an underlying SRS having phonetic speaker characteristics  
stored therein.

PRINTED IN U.S.A. 04/02/2000 10:15 AM